CELP 音声合成器とプロトタイプ分類器に基づく音声認識手法の小規模データに対する性能評価：子供音声への適用

樋口　英輔　　大崎　美穂　　白浜　公章
同志社大学大学院

1. はじめに

エンドツーエンド深層学習(E2E DL)による音声認識は，学習に大規模データを要する．モデルの複雑さゆえに，大規模データと性質が大きく異なる小規模データでは学習や転移が困難と考えられる．そこで我々は過去に，CELP 音声合成器とプロトタイプ分類器で構成される音声認識手法を提案した[1]．提案手法は，MCE 学習法によって合成と認識の両面からパラメータを最適化する．この構造と学習過程が，E2E DL では難しい小規模データに対する効率的な学習と高い説明可能性を実現し得る．

本研究では，大人音声とは性質が異なる数千規模の小規模データである子供音声を対象として，提案手法の有効性を検証する．見守りや教育における子供音声認識のニーズが高まっており，提案手法が将来的にこれらの応用に役立つことを念頭に置いている．

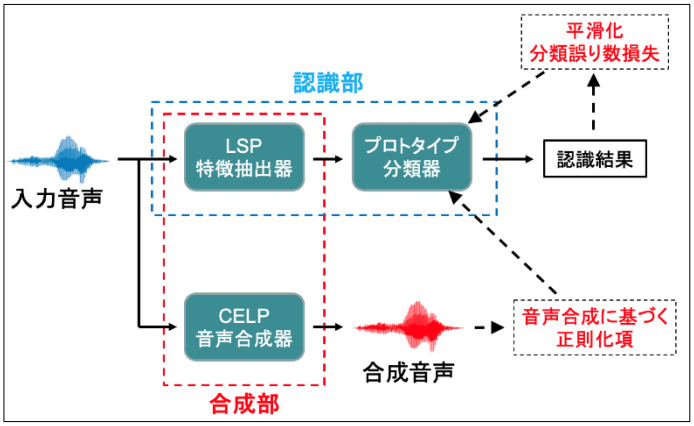
2. 提案手法

提案手法は線スペクトル対 LSP に基づく CELP 音声合成器とプロトタイプ分類器から成る(図 1 参照)．E2E DL よりもコンパクトでパラメータ数が少なく，内部の信号処理機構は説明可能である．学習では，誤分類を直接減らす MCE 学習法により，合成・認識の相互作用を通じてパラメータを最適化する．合成音声に基づく正則化項を損失関数に組み込むことで，分類器のプロトタイプが現実の音声から乖離したアーチファクトになることを防ぐ．この仕組みが小規模データに対する効率的な学習を可能にする．

3. 評価実験

大人音声に ETL-WD II，子供音声に JWC を用い，表 1 の学習条件ごとに提案手法の学習と試験を行った．学習条件のうち，No.0 では参考として大人音声の認識率を求めた．No.1(A：大人で学習，C：子供で試験)は 63.78[%]，No.3(AC：大人と子供の混在で学習，C：子供で試験)は 63.84[%]と認識率が低い．大人と子供の音声の性質が大きく異なり，子供音声に適する学習方法が必要と分かる．

図 1：提案手法の構成



一方，No.2(C：子供で学習，C：子供で試験)の 73.84[%]，および，No.4(A：大人で学習，C：子供で転移学習，C：子供で試験)の 69.46[%]は，No.1，3 よりも 5.62～10.06[%]高い．提案手法は小規模な子供音声のみでも学習や転移が可能であり，子供音声の認識率を高めると言える．

学習条件No.	学習用標本	試験用標本	認識率
0(A,A)	Adult	Adult	82.83%
1(A,C)	Adult	Child	63.78%
2(C,C)	Child	Child	73.84%
3(AC,C)	Adult + Child	Child	63.84%
4(A,C,C)	Adult → Child(**転移学習**)	Child	69.46%

表 1：実験における学習条件と認識率

4. おわりに

小規模データに対する提案手法の有効性を検証すべく，子供音声認識をタスクとする実験を行った．その結果，提案手法は子供音声のみによる学習，大人音声から子供音声への転移に有効であることが示された．今後は，子供音声や他の小規模データ(方言など)を対象として，提案手法と E2E DL 音声認識手法を比較したい．

参考文献

1) N. Umezaki et al.,: Minimum Classification Error Training with Speech Synthesis-Based Regularization for Speech Recognition, DOI:10.1145/3372806.3372819 (2019).